Journal of Computer Science and Cybernetics, V.42, N.1 (2026), 57–72
DOI: 10.15625/1813-9663/23095

ENHANCING THE HATTEN MODEL FOR LOCAL CITATION RECOMMENDATION USING BILSTM AND ATTENTION POOLING

TRAN DANG KHOA[1], THI N. DINH[2], PHU PHAM[1], BAY VO[1], NGUYEN NHU SON[3,*]

[1]*Faculty of Information Technology, HUTECH University, 475A Dien Bien Phu Street, Thanh My Tay Ward, Ho Chi Minh City, Viet Nam*
[2]*Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Street, Nghia Do Ward, Ha Noi, Viet Nam*
[3]*Institute of Information Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Street, Nghia Do Ward, Ha Noi, Viet Nam*

Abstract. Over the past decade, citation recommendation has gained increasing attention due to the exponential growth of scientific publications. Among various approaches, local citation recommendation (LCR) - a content-based method leveraging textual context - has proven effective but faces scalability challenges when applied to large databases. To balance computational efficiency and accuracy, recent systems adopt a two-stage pipeline: a lightweight prefetching phase followed by a refined reranking stage. Building upon this direction, This paper proposes Enhanced-HAtten, an improved version of the HAtten-SciBERT model [1]. The proposed model retains the original two-stage architecture but augments the prefetching phase with a Bidirectional Long Short-Term Memory (BiLSTM) layer and attention pooling, enabling richer sequential and semantic representations. Experiments on two benchmark datasets-ACL-200 and FullTextPeerRead demonstrate that Enhanced-HAtten consistently outperforms the original HAtten-SciBERT pipeline, yielding over 10% improvement in both Mean Reciprocal Rank (MRR) and Recall@K, confirming its effectiveness for large-scale scholarly recommendation tasks.

Keyword. Local citation recommendation, BiLSTM, deep learning, natural language processing, attention pooling.

1. INTRODUCTION

Citations are essential for recognizing scientific contributions and structuring connections across publications; they help interpret context and intent and allow tracing the evolution of ideas [2]. However, the exponential growth of scholarly outputs in recent decades has reshaped the research landscape. This surge causes "information overflow," overwhelming researchers with the volume of available articles. Consequently, discovering relevant findings and staying current becomes increasingly challenging. Manual review becomes inefficient

*Corresponding author.
E-mail addresses: td.khoa@hutech.edu.vn (T. D. Khoa); dinhngocthi@gmail.com (T. N. Dinh); pta.phu@hutech.edu.vn (P. Pham); vd.bay@hutech.edu.vn (B. Vo); nnson@ioit.ac.vn (N. N. Son).

and error-prone, risking omissions or reliance on less reliable sources [3, 4]. To navigate this scale, automated literature discovery and citation recommendation are indispensable. Such systems efficiently surface pertinent studies for a given context, improving research quality and efficiency [5, 6, 7]. Citation recommendation systems automatically suggest relevant papers for a given text query. These systems are categorized into global and local approaches depending on the textual scope used as input. Global citation recommendation relies on the title and abstract of a paper to retrieve broadly related works, useful for general literature surveys but lacking contextual precision. Local citation recommendation (LCR), the focus of this study, uses the textual segment around a citation placeholder optionally combined with the paper's title and abstract to identify the most appropriate reference. This approach enhances contextual relevance for specific claims or discussions. A major challenge in LCR lies in balancing accuracy and efficiency when operating over millions of papers. Complex models yield high accuracy but are computationally expensive, motivating multi-stage architectures that prefetch and rerank candidates efficiently. Among recent LCR models, the two-stage HAtten-SciBERT proposed by Gu et al. [1] remains a strong baseline. It employs a hierarchical-attention encoder to efficiently compute semantic representations for both queries and candidate papers. Although effective on benchmark datasets, its representational capacity and citation accuracy can still be improved. This study proposes Enhanced-HAtten, which augments the original architecture with a Bidirectional LSTM and attention-pooling layer to better capture contextual dependencies and emphasize informative content. The main contributions are as follows:

- Analyze strengths and weaknesses of HAtten-SciBERT [1].
- Integrate BiLSTM and attention pooling to enhance prefetching performance.
- Validate the approach on ACL-200 and FullTextPeerRead, achieving $\approx$ 10% higher MRR and Recall@K.

The remainder of this paper reviews related studies (Section 2), describes Enhanced-HAtten (Section 3), presents experiments (Section 4), and concludes (Section 5).

2. RELATED WORK

Since this study extends the HAtten-SciBERT model [1], we briefly review recent advances in local citation recommendation (LCR). Abbas et al. [3] introduced a context-aware recommender addressing cold-start issues via rhetorical zone embeddings. Celik et al. [8] proposed CiteBART, a generative BART-based model reconstructing masked citations. Khan et al. [9] applied an attention model to extract reference text from citation contexts. Roy et al. [10] developed ILCiteR, an interpretable system retrieving evidence spans to justify recommendations. Jeong et al. [11] proposed a BERT-based bi-ranker jointly encoding context and candidate papers. Zeng et al. [12] used an attention-based BiLSTM for citation worthiness prediction. Dai et al. [13] combined BERT with a heterogeneous GCN for COVID-19 recommendations. Bhowmick et al. [14] integrated co-authorship and citation history, while Yin et al. [15] introduced a scalable multi-topic LCR algorithm. Collectively, these studies illustrate the shift toward deep learning–based LCR models emphasizing semantic understanding, contextual reasoning, and scalability-motivating our Enhanced-HAtten framework.

3. ENHANCED-HATTEN MODEL DESCRIPTION

In the candidate selection stage, the HAtten model [1] encodes the citation context into a semantic vector representation for retrieving candidate papers. It employs a Transformer-based paragraph encoder and an attention-based document encoder to integrate the title, abstract, and citation context into a unified representation. However, HAtten lacks sequential inductive bias [16] and uses fixed pooling [17], limiting its ability to capture contextual order and emphasize key terms. To overcome these issues, we propose Enhanced-HAtten, which integrates a Bidirectional LSTM [18, 19]. after the Transformer encoder to model sequential dependencies and replaces fixed pooling with attention pooling [17] to highlight salient semantics. These improvements yield richer query representations and enhance overall recommendation performance.

3.1. The HAtten model and its limitations

Gu et al. [1] proposed a two-stage local citation recommendation system balancing speed and accuracy. It employs a Hierarchical-Attention text encoder (HAtten) for efficient candidate prefetching and a fine-tuned SciBERT model for reranking, achieving state-of-the-art performance on ACL-200, FullTextPeerRead, RefSeer, and arXiv. The HAtten architecture includes a paragraph encoder and a document encoder. The paragraph encoder processes three segments of each query-title, abstract, and citation context-through Transformer layers. Because self-attention lacks sequential inductive bias, positional encoding (1) [16, 17] is applied to preserve word-order information

$$\mathrm{PE}(\ \mathrm{pos},\ 2\mathrm{i}) = \sin\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d}}}}\right), \quad \mathrm{PE}(\ \mathrm{pos},\ 2\mathrm{i}+1) = \cos\left(\frac{\mathrm{pos}}{10000^{\frac{2\mathrm{i}}{\mathrm{d}}}}\right). \tag{1}$$

Each encoded paragraph produces a vector representation, forming a set of paragraph embeddings

$$d = \{e_{p_1}, e_{p_2}, \ldots, e_{p_n}\}. \tag{2}$$

To distinguish the functional role of each segment (title, abstract, or citation context), each vector is augmented with a type-specific embedding [1]

$$h_i = e_i^p + e_i^{\text{type}} \in R^d. \tag{3}$$

These vectors are processed by a document-level Transformer and aggregated through multi-head pooling to obtain a unified query representation Vq. The final vector is compared with document embeddings via cosine similarity to retrieve citation candidates [1]. Although HAtten effectively encodes citation queries, it still faces several limitations: the use of a single Transformer layer limits semantic richness [1, 20]; it lacks a sequential inductive bias needed to capture word order [16, 18]; and it inadequately emphasizes key phrases in short citation contexts [21]. This study therefore enhances the document encoder in the prefetching stage to improve contextual representation and accuracy.

3.2. Enhanced-HAtten model description

To address the limitations of the original HAtten model in the prefiltering stage, we propose Enhanced-HAtten, which improves the Document Encoder by (i) capturing sequential

dependencies among textual segments and (ii) emphasizing semantically important parts of the citation query.

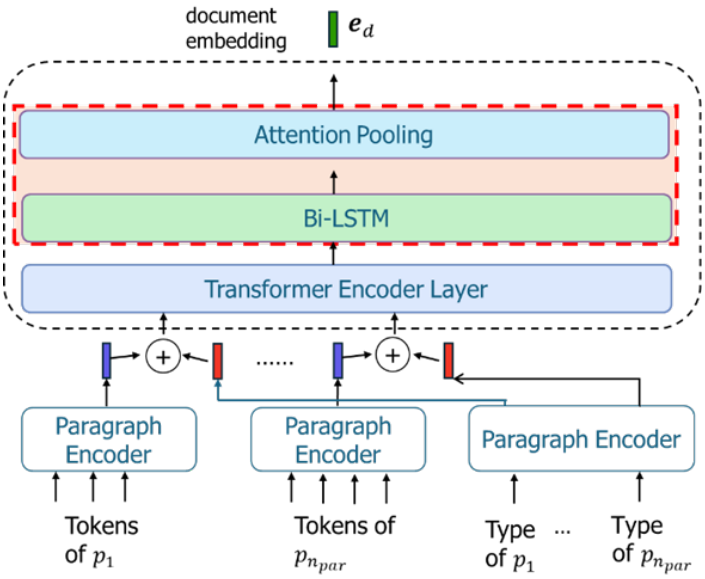


Figure 1: Enhanced document encoder architecture in the prefetching stage of Enhanced-HAtten model.

Figure 1 illustrates the enhanced document encoder architecture. While the paragraph encoder remains unchanged, the document encoder integrates a Bidirectional LSTM (BiLSTM) layer and attention pooling after the Transformer encoder. The BiLSTM models bidirectional sequential relations across paragraph embeddings, and attention pooling dynamically weights their semantic importance to obtain a more expressive representation e_d. The input to the Document Encoder comprises three paragraph embeddings from the paragraph encoder

$$d = \{e_{p_1}, e_{p_2}, e_{p_3}\}. \tag{4}$$

Each paragraph vector is enriched with a trainable type embedding that encodes its semantic role (e.g., title, abstract, or citation context), resulting in a more expressive representation. Each paragraph vector is enriched with a learnable type embedding t_i reflecting its semantic role (title, abstract, or citation context)

$$\widetilde{e_{p_i}} = e_{p_i} + t_i, \quad i = 1, 2, 3. \tag{5}$$

The resulting sequence $\{\widetilde{e_{p_1}}, \widetilde{e_{p_2}}, \widetilde{e_{p_3}}\}$ is processed by a Transformer to capture global dependencies and then passed to a BiLSTM layer that recovers sequential context. For each position t_i

$$\vec{h}_t = \text{LSTM}\left(x_t, \vec{h}_{t-1}\right), \quad \overleftarrow{h}_t = \text{LSTM}\left(x_t, \overleftarrow{h}_{t+1}\right), \quad h_t = \left[\vec{h}_t; \overleftarrow{h}_t\right]. \tag{6}$$

where x_t is the Transformer output at position t, and h_t is the concatenated forward and backward hidden states. To compress the sequence into a single query vector, attention

pooling is applied. Each hidden state h_t is projected to an intermediate score

$$u_t = \tanh\left(Wh_t + b\right). \tag{7}$$

The attention weights are computed as

$$\alpha_t = \frac{\exp(u_t^\top v)}{\sum_{j=1}^{n} \exp(u_j^\top v)}. \tag{8}$$

The final query representation is then obtained as the weighted sum

$$v_q = \sum_{t=1}^{n} \alpha_t h_t. \tag{9}$$

where W, b, and v are trainable parameters. This mechanism lets the model focus on semantically salient parts of the citation query, inspired by Attentive Pooling Networks [22]. Although Transformer self-attention effectively models long-range dependencies, it is permutation-equivariant and lacks a strong sequential inductive bias [16, 23]. Adding a BiLSTM layer restores bidirectional order awareness—essential in scholarly discourse [17, 24, 25] - while attention pooling provides adaptive emphasis on informative phrases [22, 26, 27]. Together, these components yield richer, context-sensitive embeddings, explaining Enhanced-HAtten's gains in MRR and Recall@K.

The training objective follows the triplet-loss formulation [1]

$$\mathcal{L} = \max[s(q, d^-) - s(q, d^+) + m, 0]. \tag{10}$$

where $s(q, d^-)$ denotes cosine similarity between the query and document embeddings, and m is the margin hyper-parameter fixed as in the original HAtten configuration. In summary, Enhanced-HAtten retains the Transformer's global-context modeling while introducing two crucial components - BiLSTM for sequential structure and attention pooling for adaptive focus-producing more expressive and accurate query representations for local citation recommendation.

4. EVALUATION

In this section, we begin by providing a detailed explanation of the process for constructing the Enhanced-HAtten. Next, we present the datasets and evaluation metrics utilized in our experiments. Finally, we perform a quantitative analysis comparing the Enhanced-HAtten model with its alternative approaches.

4.1. Experimental setups and implementation

Enhanced-HAtten was trained following the same pipeline and hyperparameter settings as the original HAtten [1], differing only in computing environment. We used 200-dimensional GloVe embeddings [28] as static word vectors and optimized the model with Adam [29] (learning rate = 10^{-4}, weight decay = 10^{-5}, batch size = 64). Each batch comprised one positive (cited) and five negatives, four from the top-100 BM25 results and one randomly sampled, supporting both hard-negative and soft-positive mining [1]. Training was performed on an NVIDIA RTX 3060 GPU (12 GB), with checkpoints monitored to ensure stable convergence.

4.2. Dataset description

The Enhanced-HAtten model was trained and evaluated on two standard datasets for local citation recommendation: ACL-200 and FullTextPeerRead [1, 30, 31]. ACL-200 contains about 49k citation contexts extracted from 19k ACL conference papers, with each context comprising a 200-character window around the citation marker. FullTextPeerRead, derived from the PeerRead corpus, includes 4.8k papers and 16.6k contexts, providing complete title, abstract, and citation text for both citing and cited papers. Both datasets were tokenized and normalized following the original HAtten preprocessing pipeline, enabling a consistent comparison of model generalization across citation domains.

4.3. Evaluation metrics and baseline methods

The performance of Enhanced-HAtten was evaluated under the same settings as the original HAtten framework [1], using two standard metrics: Mean Reciprocal Rank (MRR) and Recall@K (R@K), where K = 10,..., 2000. MRR reflects ranking precision, while Recall@K measures coverage across the top-K results, providing complementary insights into retrieval quality. For comparison, we adopted the same baselines as [1]: BM25 [32], Sent2Vec [33], NNSelect, and the original HAtten. In addition, ILCiteR [10] was referenced as a recent interpretable method (Recall@10 $\approx$ 0.51-0.63 on different corpora). All models were evaluated on the ACL-200 and FullTextPeerRead datasets under identical experimental conditions to ensure fairness and reproducibility.

4.4. Evaluation results

Tables 1 and 2 summarize the MRR and Recall@K (10–2000) performance of Enhanced-HAtten and baselines (BM25, Sent2Vec, NNSelect, HAtten) on the ACL-200 and FullTextPeerRead datasets.

Table 1: Performance comparison of different models on the ACL-200 dataset on Recall@K and MRR

Model	MRR	R@10	R@20	R@50	R@100
BM25	0.138	0.263	-	-	0.520
Sent2vec	0.066	0.127	-	-	0.323
NNSelect	0.076	0.150	-	-	0.402
HAtten	0.148	0.281	-	-	0.603
Enhanced - HAtten	0.167	0.305	0.396	0.522	0.621
Model	R@200	R@500	R@1000	R@2000	
BM25	0.604	0.712	0.791	0.859	
Sent2vec	0.407	0.533	0.640	0.742	
NNSelect	0.498	0.631	0.722	0.797	
HAtten	0.700	0.803	0.870	0.924	
Enhanced - HAtten	0.713	0.814	0.880	0.933	

Table 2: Performance comparison of different models on the FullTextPeerRead dataset on Recall@K and MRR

Model	MRR	R@10	R@20	R@50	R@100
BM25	0.185	0.328	-	-	0.609
Sent2vec	0.121	0.215	-	-	0.462
NNSelect	0.130	0.255	-	-	0.572
HAtten	0.167	0.306	-	-	0.649
Enhanced - HAtten	0.191	0.337	0.431	0.572	0.676
Model	R@200	R@500	R@1000	R@2000	
BM25	0.694	0.802	0.877	0.950	
Sent2vec	0.561	0.694	0.794	0.898	
NNSelect	0.672	0.790	0.869	0.941	
HAtten	0.750	0.803	0.870	0.976	
Enhanced - HAtten	0.776	0.814	0.885	0.981	

On both datasets, Enhanced-HAtten consistently surpasses baselines. On ACL-200, it achieves MRR = 0.167 and R@2000 = 0.933, higher than HAtten (0.148, 0.918) and BM25 (0.138, 0.859). On FullTextPeerRead, MRR = 0.191 and R@2000 = 0.981, it again outperforms HAtten (0.167, 0.976). These gains confirm the robustness and scalability of the proposed enhancements.

Table 3: Final ranking on FullTextPeerRead

Model	Pipeline type	R@10	MRR
BERT-GCN (Jeong et al., 2019)	Single-stage (end-to-end)	0.529	0.418
Enhanced-HAtten + Rerank (ours)	Two-stage (prefetch + rerank)	0.540	0.227

Note: BERT-GCN results are taken from [11] under full-corpus ranking, while our Enhanced-HAtten + Rerank results use the top-100 candidates, consistent with [1]. Although the pipelines differ, this setup offers a comparable assessment of final ranking effectiveness.

Table 3 compares Enhanced-HAtten + Rerank with BERT-GCN [11] under equivalent evaluation settings. Although Enhanced-HAtten achieves slightly higher Recall@10 (0.54 vs. 0.529), its MRR (0.227) remains lower than BERT-GCN (0.418) due to architectural differences: BERT-GCN integrates citation-network structure for end-to-end ranking, while Enhanced-HAtten focuses on efficient candidate prefetching followed by lightweight reranking. As confirmed by the ablation results (Sec. 4.5), BiLSTM mainly enhances MRR (fine ranking) and attention pooling improves Recall@K (coverage).

Figures 2 and 3 further illustrate that Enhanced-HAtten consistently achieves higher recall across all K values, especially at large cut-offs (R@500-R@2000). The gains stem from two architectural changes-BiLSTM capturing sequential dependencies and attention pooling emphasizing salient semantic phrases. Overall, the results demonstrate that Enhanced-HAtten significantly improves contextual citation retrieval over previous methods.

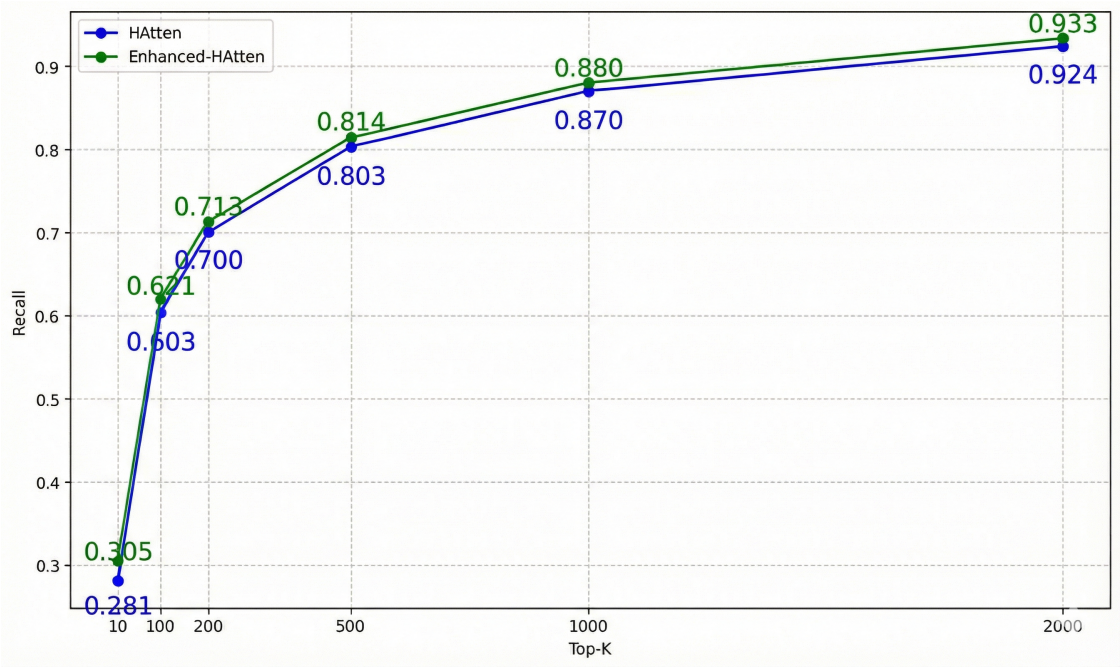


Figure 2: Recall@K comparison between HAtten and Enhanced-HAtten on the ACL-200 dataset.

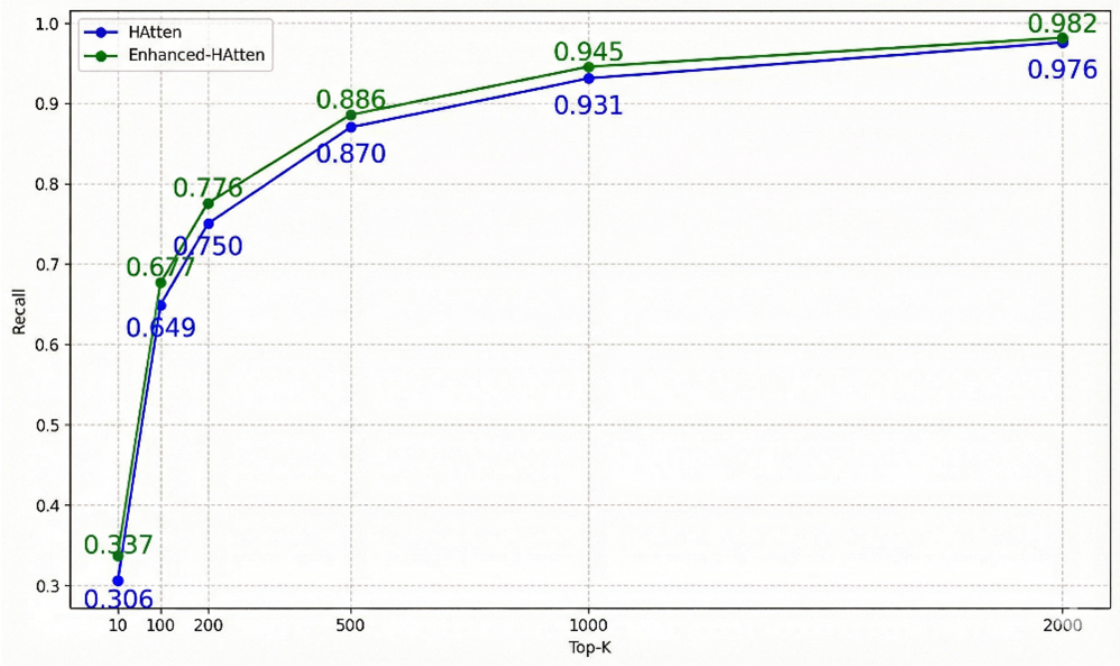


Figure 3: Recall@K comparison between HAtten and Enhanced-HAttenon the FullTextPeerRead dataset.

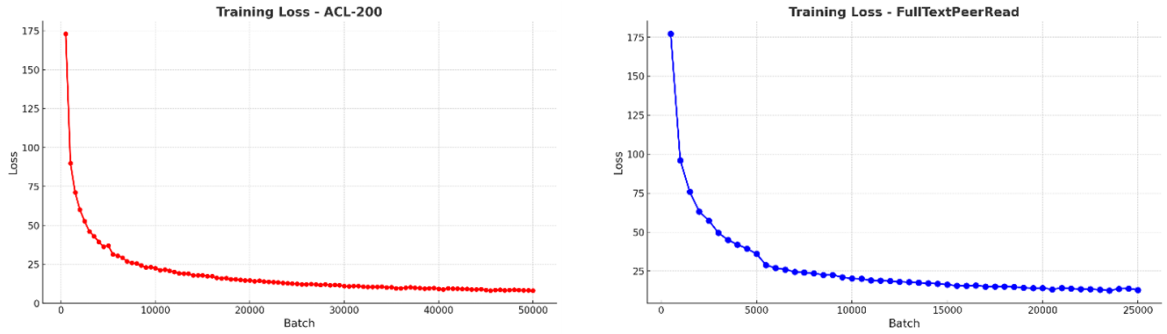
4.5. Ablation study

To evaluate the contribution of each component in Enhanced-HAtten, we examined four variants: (i) HAtten, (ii) HAtten + BiLSTM, (iii) HAtten + attention pooling, and (iv) the full model combining both. Results on ACL-200 and FullTextPeerRead are shown in Table 4.

Table 4: Results of the ablation study on ACL-200 and FullTextPeerRead

Model Variant	MRR	R@10	R@50	MRR	R@10	R@50
HAtten	0.148	0.281	-	0.167	0.306	-
+ BiLSTM	0.162	0.298	0.512	0.193	0.334	0.565
+ Attention pooling	0.164	0.301	0.517	0.183	0.332	0.577
Enhanced-HAtten	0.167	0.305	0.522	0.191	0.337	0.572

Notes that for HAtten, R@50 was not reported in the original paper and is therefore shown as "-". The results reveal complementary effects of the two modules. On ACL-200, adding BiLSTM improves MRR (0.148→0.162) and R@10 (0.281→0.298), confirming better top-rank precision. Attention pooling slightly raises R@50 (0.517 vs. 0.512), enhancing coverage through contextual aggregation. On FullTextPeerRead, BiLSTM yields the best MRR (0.193) and attention pooling achieves the highest R@50 (0.577). The combined Enhanced-HAtten attains the most balanced outcome (MRR = 0.191, R@10 = 0.337), demonstrating that BiLSTM sharpens ranking accuracy while attention pooling broadens retrieval coverage, producing a strong trade-off between precision and recall.



(a) Training loss on ACL-200 (b) Training loss on FullTextPeerRead

Figure 4: Training loss curves of Enhanced-HAtten on (a) ACL-200 and (b) FullTextPeerRead datasets, demonstrating smooth convergence and stable training behavior.

4.6. Complexity and efficiency analysis

The Enhanced-HAtten model shows stable convergence and efficient scalability on both ACL-200 and FullTextPeerRead. Each training epoch took about 33-35 minutes, with smoothly decreasing loss and no sign of overfitting (Figure 4). GPU memory usage ranged from 11–12 GB, and inference latency scaled linearly with candidate size K (0.21 s at K = 200; 0.94 s at K = 2000), confirming real-time feasibility. Theoretically, the Transformer

encoder has complexity $O(L^2d)$; the BiLSTM adds $O(Ld^2)$; and attention pooling incurs near-linear cost. These additions enhance sequential and semantic modeling with minimal overhead, maintaining an effective balance between accuracy and computational efficiency. Overall, the results confirm that Enhanced-HAtten achieves efficient, stable training and low-latency inference, supporting its practical applicability to large-scale citation recommendation systems.

4.7. Case study and error analysis

To complement the quantitative evaluation, we conducted a qualitative case study to analyze how Enhanced-HAtten behaves compared to the original HAtten. Table 5 presents representative citation contexts from ACL-200 and FullTextPeerRead, showing the ground-truth citation and the top-10 candidates retrieved by both models. (Note: paper names listed are ground-truth labels from the benchmark datasets rather than new references.)

Successful case: In context #3490 (ACL-200), the target citation "Generalizing Word Lattice Translation (2008)" was ranked 1st by Enhanced-HAtten but missed entirely by HAtten. This demonstrates how sequential inductive bias (BiLSTM) and phrase-level salience (attention pooling) help capture key lexical cues such as "word lattice."

Partially improved case: In context #718 (ACL-200), "Extracting Lexically Divergent Paraphrases from Twitter (2014)" was ranked 5th by Enhanced-HAtten and 94th by HAtten. Although multiple paraphrase-related papers caused confusion, the enhanced model still achieved substantial improvement.

Failure case: In short or ambiguous contexts, both models sometimes failed to retrieve the correct citation within the top-10, reflecting the difficulty of identifying targets without distinctive lexical anchors.

Table 5: Case study: comparison between Enhanced-HAtten and HAtten on representative citation contexts from ACL-200 and FullTextPeerRead

Context Id	Context	Ground truth	Model	Dataset	Predicted Citations (Top 10 candidates)
3490	... decoding process of SMT <TARGETCIT >exponential number of alternatives ...	Dyer et al., 2008	Enhanced-HAtten	ACL-200	1. Dyer et al., 2008 – Generalizing Word Lattice Translation. 2. Jiang et al., 2011 – Incorporating Source-Language Paraphrases... 3. Li et al., 2009 – Collaborative Decoding: Partial Hypothesis... 4. Onishi et al., 2010 – Paraphrase Instance for SMT 5. Feng et al., 2009 – Lattice-based System Combination for SMT 6. Schroeder et al., 2009 – Word Lattices for Multi-Source... 7. Federmann et al., 2009 – Translation Combination using... 8. ACL Moy-1046 – Using a Maximum Entropy Model. 9. Aziz et al., 2014 – Exact Decoding for PB-SMT 10. Durrani et al., 2013 – Model With Minimal Translation Units...

Context Id	Context	Ground truth	Model	Dataset	Predicted Citations (Top 10 candidates)
3490	... decoding process of SMT <TARGETCIT >exponential number of alternatives ...	Dyer et al., 2008	HAtten	ACL-200	1. Koehn et al., 2015 – Extended Translation Models... 2. Hu et al., 2015 – Context-Dependent Translation... 3. Chiang et al., 2010 – Assessing Phrase-Based Translation.. 4. Yamada & Knight, 2003 – Greedy Decoding for SMT... 5. Rosti et al., 2010 – Boosting-based System Combination for MT 6. Hardmeier et al., 2012 – Document-Wide Decoding for PB-SMT 7. Takezawa et al., 2007 – NICT-ATR Speech-to-Speech... 8. Zhao et al., 2011 – Hypothesis Mixture Decoding for SMT 9. Durrani et al., 2015 – The Operation Sequence Model... 10. Hardmeier et al., 2013 – Docent: A Document-Level... (Ground truth is outside the list of 100 candidates)
718	... DLS@CU system (OTHERCIT) ... MULTIP latent variables model <TARGETCIT >utilizes anchor pairs ...	Xu et al., 2014	Enhanced-HAtten	ACL-200	1. Mimo et al., 2010 – Cross-Lingual Latent Topic Extraction... 2. Xu et al., 2015 – TKLBLR: Detecting Twitter Paraphrases... 3. Das & Smith, 2009 – Paraphrase Identification as Probabilistic... 4. Zhao et al., 2006 – BitAM: Bilingual Topic AdMixture Models... 5. Xu et al., 2014 – Extracting Lexically Divergent Paraphrases from Twitter... 6. Socher et al., 2012 – Modeling Sentences in the Least Paraphrases... 7. Turney, 2006 – Latent Variable Models for Semantic... 8. Gruber et al., 2008 – Latent Variable Models for... 9. Tamura et al., 2011 – Identifying Word Translations... 10. Quirk et al., 2013 – Semi-Word-Phrase-Based...
718	... DLS@CU system (OTHERCIT) ... MULTIP latent variables model <TARGETCIT >utilizes anchor pairs ...	Xu et al., 2014	HAtten	ACL-200	1. Das & Smith, 2009 – Paraphrase Identification as Probabilistic.. 2. Liu et al., 2013 – A Lightweight and High Performance.. 3. Quirk et al., 2004 – Monolingual Machine Translation.. 4. Bannard & Callison-Burch, 2005 – Paraphrasing... 5. Wan et al., 2006 – Paraphrase Recognition via... 6. MacCartney & Manning, 2008 – A Phrase-Based Alignment.. 7. Tiedemann, 2009 – Collocation Extraction... 8. Xu et al., 2015 – TKLBLIIR: Detecting Twitter Paraphrases... 9. Barzilay & Lee, 2004 – A Phrase-Based HMM Approach... 10. Ji et al., 2013 – Paraphrasing Adaptation... (Ground truth is ranked at #94, not in the top-10)

Context Id	Context	Ground truth	Model	Dataset	Predicted Citations (Top 10 candidates)
1331	... experience replay ... can be preferentially sampled <TARGETCIT >... balances reward across trajectories ...	Schaul et al., 2015	Enhanced-HAtten	FullTextPeerRead	1. Osband et al., 2016 – Deep Exploration via Bootstrapped DQN 2. Schaul et al., 2015 – Prioritized Experience Replay 3. Bellemare et al., 2016 – A Study of Count-Based Exploration... 4. Lakshminarayanan et al., 2016 – Deep Reinforcement Learning.. 5. van Hasselt et al., 2017 – Learning to Repeat: Fine-Grained.. 6. Liang et al., 2017 – Fine-Grained Acceleration Control.. 7. Rusu et al., 2015 – Policy Distillation 8. Oh et al., 2016 – Dynamic Frame Skip Deep Q-Network 9. Gruslys et al., 2017 – The Reactor: A Sample-Efficient.. 10. Isele et al., 2017 – Navigating Intersections with Autonomous..
1331	... experience replay ... can be preferentially sampled <TARGETCIT >... balances reward across trajectories ...	Schaul et al., 2015	Enhanced-HAtten	FullTextPeerRead	1. Isele et al., 2017 – Navigating Intersections.. 2. Wei et al., 2017 – Traffic Light Control Using Deep.. 3. Rusu et al., 2016 – Multi-task Learning with.. 4. Gu et al., 2017 – Learning Control for Air Hockey.. 5. Hwangbo et al., 2017 – Autonomous Quadrotor Landing... 6. Wang et al., 2015 – Dueling Network Architectures.. 7. Mnih et al., 2013 – Playing Atari with Deep RL 8. Tai et al., 2017 – Virtual-to-Real DRL for Mapless.. 9. Lakshminarayanan et al., 2016 – Deep RL With... 10. Arulkumaran et al., 2017 – Deep Reinforcement.. (Ground truth is ranked at #36, not in the top-10)

Overall, Enhanced-HAtten not only improves aggregate metrics (MRR, Recall@K) but also elevates the correct citation's rank in individual contexts. The remaining errors highlight opportunities for integrating richer contextual signals and advanced negative sampling strategies in future work.

Table 6: Quantitative error analysis of HAtten and Enhanced-HAtten (prefetch stage)

Error type	PeerRead		ACL-200	
	HAtten	Enhanced-HAtten	HAtten	Enhanced-HAtten
Type-A (Miss@K)	2.26% (154)	2.10% (143)	7.27% (697)	6.66% (638)
Type-B (Near-miss)	66.03% (4499)	64.51% (4396)	64.62% (6194)	63.79% (6114)
Type-C (Drift)	23.48% (1600)	25.17% (1715)	21.26% (2038)	24.34% (2333)
Type-D (On-topic wrong)	0.19% (13)	0.25% (17)	0.11% (11)	0.08% (8)
OK (Top-1)	8.04% (548)	7.97% (543)	6.73% (645)	5.13% (492)
Total	6814	6814	9585	9585

Compared with the original HAtten, Enhanced-HAtten markedly reduces Type-A errors, confirming stronger candidate coverage, though Type-C (Drift) slightly increases and OK

cases decrease marginally. This trade-off indicates that while the model rarely misses the correct citation, it sometimes ranks a thematically similar paper first. Such behavior explains its higher Recall@K but slightly lower MRR than graph-aware reranking models like BERT-GCN, reflecting the intended balance between broad coverage and ranking precision.

5. CONCLUSION AND FUTURE WORK

This study proposed Enhanced-HAtten, an improved model for local citation recommendation at the prefetching stage. By extending the original Document Encoder with a BiLSTM layer to capture sequential dependencies and applying attention pooling to highlight salient information, the model overcomes the limitations of Transformer-based architectures-namely, the lack of sequential inductive bias and reliance on static pooling [1, 17].

Experiments on two benchmark datasets, ACL-200 and FullTextPeerRead, demonstrate that Enhanced-HAtten consistently outperforms both the original HAtten and other baselines, achieving 13-14% improvement in MRR and 3-6% gain in Recall@K ($K \geq 50$), validating the effectiveness of the proposed modifications.

Future work will explore: (i) extending the model to multilingual contexts, including Vietnamese, using NLP toolkits such as VnCoreNLP [34]; (ii) improving scalability through approximate nearest-neighbor search (e.g., FAISS, HNSW) [35]; and (iii) enhancing interpretability via explainable-AI techniques like SHAP[1] [36]. Practical deployment may also integrate the model with academic platforms such as Google Scholar[2] , Semantic Scholar[3] , or Scite.ai[4] [37] to support researchers in efficiently retrieving relevant citations.

ACKNOWLEDGEMENT

This work was supported by Vietnam Academy of Science and Technology, under Grant VAST01.03/25-26.

REFERENCES

[1] N. Gu, Y. Gao, and R. H. R. Hahnloser, "Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking," in *European Conference on Information Retrieval (ECIR)*, 2022. [Online]. Available: https://arxiv.org/abs/2112.01206

[2] Y. Zhang, Y. Wang, Q. Z. Sheng, L. Yao, H. Chen, K. Wang, A. Mahmood, W. E. Zhang, M. Zaib, S. Sagar, and R. Zhao, "Deep learning meets bibliometrics: A survey of citation function," *Journal of Informetrics*, 2025. [Online]. Available: https://doi.org/10.1016/j.joi.2024.101608

[3] M. A. Abbas, S. Ajayi, M. Bilal, A. Oyegoke, M. Pasha, and H. T. Ali, "A deep learning approach for context-aware citation recommendation using rhetorical zone classification and similarity to overcome cold-start problem," *Journal of Ambient Intelligence and Humanized Computing*, vol. 15, pp. 419–433, 2022. [Online]. Available: https://doi.org/10.1007/s12652-022-03899-6

[1] https://shap.readthedocs.io/
[2] https://scholar.google.com/
[3] https://www.semanticscholar.org/
[4] https://scite.ai/

[4] Y. Lu, M. Yuan, J. Liu, and M. Chen, "Research on semantic representation and citation recommendation of scientific papers with multiple semantics fusion," *Scientometrics*, vol. 128, pp. 1367–1393, 2023. [Online]. Available: https://doi.org/10.1007/s11192-022-04566-5

[5] Y. Liang and L. K. Lee, "A systematic review of citation recommendation over the past two decades," *International Journal on Semantic Web and Information Systems*, vol. 19, no. 1, pp. 1–22, 2023. [Online]. Available: https://doi.org/10.4018/IJSWIS.324071

[6] Z. Medić and J. Šnajder, "A survey of citation recommendation tasks and methods," *Journal of Computing and Information Technology*, vol. 28, no. 3, pp. 183–205, 2020. [Online]. Available: https://doi.org/10.20532/cit.2020.1005160

[7] Z. Ali, I. Ullah, A. Khan, A. Ullah Jan, and K. Muhammad, "An overview and evaluation of citation recommendation models," *Scientometrics*, vol. 126, no. 5, p. 4083–4119, 2021. [Online]. Available: http://dx.doi.org/10.1007/s11192-021-03909-y

[8] E. Y. Çelik and S. Tekir, "Citebart: Learning to generate citations for local citation recommendation," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, p. 1703–1719. [Online]. Available: http://dx.doi.org/10.18653/v1/2025.emnlp-main.89

[9] D. Khan, I. Ahmed, I. Ullah, and A. Alwabli, "Finding the reference text in citation contexts using attention model," *Service Oriented Computing and Applications*, vol. 19, no. 1, p. 45–55, 2024. [Online]. Available: http://dx.doi.org/10.1007/s11761-024-00410-1

[10] S. G. Roy and J. Han, "Ilciter: Evidence-grounded interpretable local citation recommendation," 2024. [Online]. Available: https://arxiv.org/abs/2403.08737

[11] C. Jeong, S. Jang, E. Park, and S. Choi, "A context-aware citation recommendation model with bert and graph convolutional networks," *Scientometrics*, vol. 124, no. 3, p. 1907–1922, 2020. [Online]. Available: http://dx.doi.org/10.1007/s11192-020-03561-y

[12] T. Zeng and D. E. Acuna, "Modeling citation worthiness by using attention-based bidirectional long short-term memory networks and interpretable models," *Scientometrics*, vol. 124, no. 1, p. 399–428, 2020. [Online]. Available: http://dx.doi.org/10.1007/s11192-020-03421-9

[13] T. Dai, J. Zhao, D. Li, S. Tian, X. Zhao, and S. Pan, "Heterogeneous deep graph convolutional network with citation relational bert for covid-19 inline citation recommendation," *Expert Systems with Applications*, vol. 213, p. 118841, 2023. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2022.118841

[14] A. Bhowmick, A. Singhal, and S. Wang, "Augmenting context-aware citation recommendations with citation and co-authorship history," in *18th International Conference on Scientometrics and Informetrics, ISSI 2021*, 2021, pp. 115–120.

[15] M. J. Yin, B. Wang, and C. Ling, "A fast local citation recommendation algorithm scalable to multi-topics," *Expert Systems with Applications*, vol. 238, p. 122031, 2024. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2023.122031

[16] M. Hahn, "Theoretical limitations of self-attention in neural sequence models," *Transactions of the Association for Computational Linguistics*, vol. 8, p. 156–171, 2020. [Online]. Available: http://dx.doi.org/10.1162/tacl_a_00306

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[18] I. Sonata and Y. Heryadi, "Comparison of lstm and transformer for time series data forecasting," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, 2024, p. 491–495. [Online]. Available: http://dx.doi.org/10.1109/ICICoS62600.2024.10636892

[19] D. Purwitasari, A. F. Abdillah, S. Juanita, I. K. E. Purnama, and M. H. Purnomo, "A comparison of transformer and bilstm based bioner model with self-training on low-resource language texts of online health consultation," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 6, pp. 213–224, 2023. [Online]. Available: https://doi.org/10.22266/ijies2023.1231.18

[20] Z. Medić and J. Šnajder, "An empirical study of the design choices for local citation recommendation systems," *Expert Systems with Applications*, vol. 200, p. 116852, 2022. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2022.116852

[21] Y. Zhang and Q. Ma, "Dual attention model for citation recommendation," 2020. [Online]. Available: https://arxiv.org/abs/2010.00182

[22] C. d. Santos, M. Tan, B. Xiang, and B. Zhou, "Attentive pooling networks," 2016. [Online]. Available: https://arxiv.org/abs/1602.03609

[23] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1906.08237

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," p. 4171–4186, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/N19-1423

[25] D. Oralbekova, O. Mamyrbayev, S. Zhumagulova, and N. Zhumazhan, "A comparative analysis of lstm and bert models for named entity recognition in kazakh language: A multi-classification approach," p. 116–128, 2024. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-72260-8_10

[26] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional lstm with two-dimensional max pooling," 2016. [Online]. Available: https://arxiv.org/abs/1611.06639

[27] P. Maini, K. Kolluru, D. Pruthi, and Mausam, "Why and when should you pool? analyzing pooling in recurrent architectures," 2020. [Online]. Available: https://arxiv.org/abs/2005.00159

[28] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, p. 1532–1543. [Online]. Available: http://dx.doi.org/10.3115/v1/D14-1162

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[30] D. Kang, W. Ammar, B. Dalvi, M. van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz, "A dataset of peer reviews (peerread): Collection, insights and nlp applications," 2018. [Online]. Available: https://arxiv.org/abs/1804.09635

[31] Z. Medić and J. Snajder, "Improved local citation recommendation based on context enhanced with global information," in *Proceedings of the First Workshop on Scholarly Document Processing*, 2020, p. 97–103. [Online]. Available: http://dx.doi.org/10.18653/v1/2020.sdp-1.11

[32] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, p. 333–389, 2009. [Online]. Available: http://dx.doi.org/10.1561/1500000019

[33] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018. [Online]. Available: http://dx.doi.org/10.18653/v1/N18-1049

[34] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "Vncorenlp: A vietnamese natural language processing toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2018. [Online]. Available: http://dx.doi.org/10.18653/v1/N18-5012

[35] J. Johnson, M. Douze, and H. Jegou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, p. 535–547, 2021. [Online]. Available: http://dx.doi.org/10.1109/TBDATA.2019.2921572

[36] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[37] J. M. Nicholson, M. Mordaunt, P. Lopez, A. Uppala, D. Rosati, N. P. Rodrigues, P. Grabitz, and S. C. Rife, "scite: A smart citation index that displays the context of citations and classifies their intent using deep learning," *Quantitative Science Studies*, vol. 2, no. 3, p. 882–898, 2021. [Online]. Available: http://dx.doi.org/10.1162/qss_a_00146

Received on June 29, 2025
Accepted on October 13, 2025